Next Generation Infrastructure for High Performance Global Big Data Science

Joe Mambretti
Director, International Center for Advanced Internet Research
Northwestern University

Increasingly, science requires the gathering, transport, analysis and storage of extremely large volumes of data, a key resource for new knowledge discovery. High energy physics, bioinformatics, genomics, oceanography, atmospheric sciences, weather modeling, astrophysics, geophysical sciences, earth sciences, materials science, chemistry, molecular dynamics, neurosciences, and many other disciplines use petabytes of information for sophisticated analytic processing, models, and simulations. Also, many of these science communities are creating next generation instruments that will generate many additional petabytes of data (exebytes in aggregate), including high luminosity synchrotrons, including the High Luminosity Large Hadron Collider, the Large Synoptic Survey Telescope, the Square Kilometer Array, neutrino detectors, genomic sequencers and others.

In general, data-intensive scientific workflows require data transport among generation/instrumentation sites, analytic/compute facilities (HPC centers, grid facilities, science clouds, analytic centers, etc.), storage/repository centers and visualization facilities. Supporting large scale, high capacity E2E data flows among such sites, especially over large distances around the world, requires specialized architecture, services, capabilities, and technologies.

In response to these requirements, research communities are designing, developing, implementing, and operating large scale distributed science environments, many building on the Science DMZ concepts created by the Energy Science Network (ESnet). These specialized, segmented environments consist of multiple instruments and facilities, interconnected by 100 Gbps high performance paths and high performance transport devices, including at edge sites. For example, open exchange sites, such as the StarLight International/National Communications Exchange Facility has 60 100 Gbps paths, with another 25 expected by the end of the year. The facility is currently experimenting and demonstrating 400 Gbps E2E and Tbps services, foundation capabilities, and technology.

One key focus of these efforts are activities creating next generation data transport services, protocols, and infrastructure, such as Data Transfer Nodes (DTNs). Other focal efforts are developing techniques based on Software Defined Networking (SDN), Software Defined Exchanges (SDXs), and Software Defined Infrastructure (SDI). Complementary projects are creating new methods for integrating large scale

flows directly with edge compute facilities, including HPC facilities and science clouds, and storage facilities, for example, the Open Storage Network, an NSF funded distributed national storage substrate.

The continuing growth in Big Data for science is fundamentally transforming environments used to transform that data into it knowledge. Currently, science communities are planning for a wide range of future instruments that will exponentially increase data production and accelerate this process. New types of distributed environments are being designed to address the challenges of this approaching data tsunami. One such environment is the Global Research Platform initiative. Many of the innovations being developed through this process will migrate to wider communities beyond science communities.